<<数据挖掘实验/统计实验教材系列>>

图书基本信息

书名：<<数据挖掘实验/统计实验教材系列>>

13位ISBN编号：9787503763649

10位ISBN编号：7503763647

出版时间：2011-9

出版时间：孔志周、肖百龙、 许涤龙 中国统计出版社 (2011-09出版)

作者：孔志周，肖百龙 著

页数：235

内容概要

《统计实验教材系列：数据挖掘实验》作为《数据挖掘》课程的实验教材，其编写的目的在于，在讲授数据挖掘基础理论和基本方法的同时，一方面，通过前十一个实用性实验项目，使学生加深对数据挖掘内涵的理解，学会利用SAS/EM进行数据挖掘的技能；另一方面，通过最后一个研究性实验项目，学习数据挖掘研究的思路，认识数据挖掘研究的迫切性，增强学生的基本研究技能，为学生工作或研究生阶段打下基础，提高运用数据挖掘相关理论进行理论分析与解决实际问题的能力。

<<数据挖掘实验/统计实验教材系列>>

书籍目录

章节摘录

版权页： 插图： 聚类分析是一种流行的数据离散化方法。
通过将属性的值划分成簇或组，聚类算法可以用来离散化数值属性。
聚类考虑属性的分布以及数据点的邻近性，因此可以产生高质量的离散化结果。
每一个簇形成概念分层的一个节点，而所有的节点在同一个概念层。
每一个簇可以进一步分成若干个子簇，形成较低的概念层。
簇也可以聚集在一起，以形成分层结构中较高的概念层。
（6）通过直观划分离散化 许多用户希望看到数值区域被划分为相对一致的、易于阅读的、看上去直观或“自然”的区间。
3—4—5规则可以用于将数值数据划分成相对一致和“自然”的区间。
一般地，该规则根据最高有效位的取值范围，递归地和逐层地将给定的数据区域划分为3、4或5个相对等宽的区间。
该规则可以递归地用于每个区间，为给定的数值属性创建概念分层。
由于在数据集中可能包含特别大的正或负的离群值，最高层分段简单地按最小或最大值可能导致扭曲的结果。
这样，顶层离散化可以根据代表给定数据大多数的数据区间（例如，第5个百分位数到第95个百分位数）进行。
越出顶层分段的特别高和特别低的值将用类似的方法形成单独的区间。
5.2.2离散数据的概化 离散数据具有有限个（但可能很多）不同值，值之间无序，例如地理位置、工作分类和商品类型。
有很多方法产生分类数据的概念分层。
（1）由用户或专家在模式级显式地说明属性的部分序：通常，分类属性或维的概念分层涉及一组属性。
用户或专家在模式级通过说明属性的部分序或全序，可以很容易地定义概念分层。
（2）通过显式数据分组说明分层结构的一部分：这基本上是人工地定义概念分层结构的一部分。
在大型数据库中，通过显式的值枚举定义整个概念分层是不现实的。
然而，对于一小部分中间层数据，可以很容易地显式说明分组。
（3）说明属性集但不说明它们的偏序：用户可以说明一个属性集形成概念分层，但并不显式地说明它们的偏序。
然后，系统可以尝试自动地产生属性的序，构造有意义的概念分层。
由于一个较高层的概念通常包含若干从属的较低层概念，定义在高概念层的属性与定义在较低概念层的属性相比，通常包含较少数目的不同值。
根据这一事实，可以根据给定属性集中每个属性不同值的个数自动地产生概念分层。
具有最多个不同值的属性放在分层结构的最底层。

编辑推荐

版权说明